Detecting Environmental Infractions and their Impacts Caused by Industrial Sectors

Anonymous ACL submission

Abstract

Environmental practices of an organization reflects its commitments to the world environment, and societal good. Institutional investors take regulatory violations into account for decision making purposes, since these factors are known to affect public opinion and thereby the stock indices of companies. Typically, risk scores are derived based on information published in the reports filed by companies, News articles and social media posts, analyst reactions along with customized surveys. Though this involves churning large volumes of textual information, not much use of language technologies is reported by practitioners for information extraction and classification for detecting environmental violations by organizations. In this paper, we present a transformer based multi-task network to help detect environmental violations from Online News articles and classify them into respective environmental impacts. We have created an annotated corpus using articles published over last 8 years, mostly by regulatory and governing agencies across different countries, for the purpose. Due to the paucity of data, we have adopted an active learning framework. We observed the models to performs better at each round when new, clean human annotations are added. Both the incident classification and extraction methods achieve state-of-the-art accuracy, as measured using cross-validation techniques.

1 Introduction

As awareness about environmental sustainability is gaining grounds across all sections of society, it is becoming increasingly clear that better sustainability practices can drive better investment outcomes too (Tarmuji et al., 2016; Shahi et al., 2014; Zhao et al., 2018; Yoon et al., 2018; kpm, 2017; Velte, 2017; Sultana et al., 2018; Raman et al., 2020). Studies (ccb, 2018) indicate that around 69% of the companies that experienced a high or severe environmental incidents, experienced an average market cap decline of 6% within the next ten days.

Assessing environmental sustainability practices involves analyzing large volumes of textual content gathered from reports published by organizations, reports published by global and regional monitoring organizations, news reports and social media content (Guo, 2020; Pasch and Ehnes, 2022; Murakami and Muraoka, 2022; Pasch and Ehnes, 2022; Liu et al., 2019; Collobert et al., 2011; Liu et al., 2015; Luong et al., 2015; Xu et al., 2018; Wan et al., 2021). Presently, most analysts collect relevant data from management personnel using surveys. Different agencies use different survey questionnaire which reportedly leads to survey fatigue for managers (ccb, 2018). Application of language technologies can enable continuous monitoring in an automated manner (Goel et al., 2020; Lee and Kim, 2023; Nugent et al., 2020).

In this paper, we present the concept of an adverse media screening framework that can provide insights about an organization's environment-related violation incidents, if any. The system utilizes an active learning based multi-task neural models to detect and classify environmental violations and their potential impacts on the earth including human health hazards, wild life, aquatic life air quality and soil.

The contributions of the paper are summarized as follows:

1. We present the concept of an adverse media screening framework that can provide insights about an organization's environment-related violation incidents.

2. Looking into the paucity of annotated data we propose to use an active learning based multi-tasking neural architecture for detecting environmental incidents, including violation clauses and their potential impacts.

3. An annotated corpus of regulatory articles to mark risks or incidents, regulatory violations and penalties along with the target organization which was reported.

1.1 Annotating Environmental Violations and Impacts

Incidents that violate environmental clauses are reported as events where the actors are the violating organizations. These reports contain the following information:

a) **Target Organization (TO)**: Of the many organization names that may appear in a document, the task during annotation is to identify and tag the violating or the award-winning organization.

b) **Environmental violations (V):** these are phrases or sets of words that collectively indicate non-compliance or failure to comply with guidelines or regulations.

c) **Potential Environmental Impact (I):** caused due to a specific environmental violation. This includes Human health, impact on soil and natural resources, aquatic life and wild life.

d) **Action taken (A):** The currency value that denotes the penalty that has been enforced upon the target organization by a governing body.

We have collected a total of 3100 documents from United States Environmental Protection Agency (US-EPA)[1], Oregon Department of Environmental Quality (DEQ)[2], and EPA-Canada[3]. All the News are published between the time-period of 2015-2023. The average length of a document is around 23 sentences. Six annotators took part in the annotation, with each expert annotating around 600 documents using the Stanford simple manual annotation tool [4]. This included 100 documents, which were sent to all the annotators to compute an inter-annotator agreement. Each document is first processed using the Stanford NER (Manning et al., 2014) to obtain the organization names, locations and currency values as named entities. The experts read each document and performed the following tasks,

Task-1: - Mark phrases in the text that indicate environmental violation.

Task-2: Label each document into any one of the four environmental impact categories: a) *Human health (H)*, b) *impact on soil and natural resources (S)*, c) *aquatic life (A)* and d) *wild life (W)*.

Using the annotations obtained for 100 common documents, we measured the inter-annotator agreement using the Fleiss Kappa (Fleiss et al., 1981) measure (κ). This is computed as $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. The factor $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. It was observed that the inter-annotator score for Task-1 was 0.83, which is appreciably high. For Task-2, it was found to be 0.71. The scores are computed using word-label matches assigned by different annotators. The very high scores indicate that all experts were marking fairly uniformly and therefore, the expert annotated dataset is reliable to be used for training incident detection systems.

Altogether, we obtained 3671 violation phrases, 2100 penalty phrases, 2223 Target Organizations, 3300 locations, and 2995 Environmental Impacts. The entire corpus can be publicly released.

2 A Multi-tasking Neural Model for Impact Classification and Violation Detection

Multi-task learning utilizes the correlation between related tasks to improve classification by learning tasks in parallel. In the present work, the two related tasks are *task-1:* classifying a document into any one of the four environmental impacts and *task-2:* labeling appropriate phrases in the text as for violation detection. It is worth mentioning here that identification of the violation phrases can be considered as a kind of explanation for the task-1 classification task ().

The proposed multi-task network uses a cascaded CNN-BiLSTM layer for the combined tasks of classification and extraction, using the fine-tuned BERT for creating the sequence embeddings.

To obtain the multi-tasking model for dual tasks of classification and extraction, the $BERT - CNN - BiLSTM$ layers have been trained with two separate loss functions L_1 and L_2. Where, $\mathrm{L}_1(\theta) = -\sum_{t=1}^{M}\sum_{k=1}^{K} \bar{y}_t^k log(y_t)$ and $L_2(\theta) = -\sum_{t=1}^{N}\sum_{j=1}^{J} \bar{q}_t^{i,j} log(q_t^i)$ q_t is the vector representation of the predicted output of the model for the input word w_t^i. K and J are the number of class labels for each task. The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

[1] https://www.epa.gov/newsreleases

[2] https://www.oregon.gov/deq/pages/enforcement-actions.aspx

[3] https://www.canada.ca/en/environment-climate-change/services/environmental-enforcement/notifications.html

[4] https://nlp.stanford.edu/software/

Table 1: Sample enforcement News with the respective annotated entities and events. Note that all the target organization names were intentionally masked by the token [ORGName] to maintain anonymity.

News	Impact
On April 26, 2024, in the Provincial Court of Newfoundland and Labrador, {**ORGName**}**TO** was ordered to pay {**$2 million**}**P** after earlier pleading guilty to one charge under the federal Fisheries Act and one charge under the Migratory Birds Convention Act, 1994.... The charges relate to a crude oil release on November 16, 2018, at the White Rose oil field in the Newfoundland and Labrador offshore area, where an **estimated 250,000 litres of crude oil were released into the environment due to a failure of the subsea flowline connector from the SeaRose Floating Production, Storage and Offloading installation.**}**V** Crude oil is deleterious to fish and harmful to migratory birds. Between November 18 and 23, 2018, 17 potentially oiled birds were observed from offshore vessels and platforms, seven of which were captured. An oiled bird was also discovered on December 4, 2018. These observations, and subsequent laboratory analyses, confirmed that the oil release affected various migratory birds.	Aquatic life

We define the joint loss function using a linear combination of the loss functions of the two tasks as: $L_{joint}(\theta) = \lambda * L_1(\theta) + (1-\lambda) * I_{[y_{sentence}=1]} * L_2(\theta)$ Where, λ controls the contribution of losses of the individual tasks in the overall joint loss. $I_{[y_{sentence}=1]}$ is an indicator function which activates the loss only when the corresponding sentence classification label is 1, since we do not want to back-propagate sequence labeling loss when the corresponding sequence classification label is 0.

2.1 Active Learning

In this section, we study the effectiveness of classification and extraction process under an active learning setup. Given a set of T documents about an environmental event and labeled with different environmental impacts, but without identified violations, we randomly select k% or S number of documents and ask human for violation annotations. Then, the labeled data is employed to train our multi-task network. Next, we use the trained violation extractor to predict violations on the unlabeled data, assign a score to every unlabeled document and select the same S number of documents with the highest scores. These documents are again given to humans for annotation and adding to the next training round. The selection of new data and training process can be stopped after k interactions or when the model performance is not significantly improved. The document score at each round is computed based on the predicted violation tokens as follows: $Score_t = 1 - \frac{\sum_{i=1}^{n}(\bar{y}_i * p_i)}{\sum_{i=1}^{n}(y)_i}$ where $\bar{y}_i \in 0, 1and$p$_i$ are the predicted label and the probability of a token w_i to be a violation phrase. n is the number of tokens in the document T.

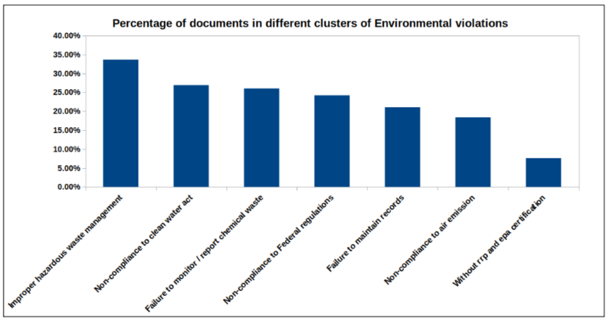

Figure 1: Distribution of violation clusters in environmental sector.

3 Evaluation

The performance of the proposed model has been compared with a number of baseline models used for single-objective document classification and sequence labeling tasks as well as large language models like LLAMA-2 7B and fined-tuned Mistral 7B, as depicted in Table 2.

Our preliminary investigation shows for almost all the categories the Multi-task BERT-CNN-BiLSTM model significantly outperforms the baseline models including LLAMA-2 and Mistral. For example, in the *Target Organization* class, it was found that the Multi-task BERT-CNN-BiLSTM model significantly reduces the false negative score and achieved a high true positive score thereby achieving a high precision and recall. In general, an F-Measure of 0.89 with a precision of 0.87 and recall of 0.92 was achieved. For the *violation* class F1 score of 0.87 with a high recall of 0.92 was obtained. However for the *Action taken* class we observed that the single task Fined-tuned Mistral7B performs better than the proposed network. Although, for both the cases the recall values are same, mistral7B classification produces a better precision of 0.86 as compared to 0.80 in our model. The target organizations were detected correctly 89% of the times. In the remaining cases, either wrong organizations were detected or missed out altogether. Detailed analysis reveals that for violation and incident phrases majority of the sub-sequences

Table 2: Results reporting the sequence classification and sequence labeling experiments

	Document			Extraction											
	Classification			TO			I			V			A		
	P	R	F1	P	R	F	P	R	F	P	R	F	P	R	F
Single task CNN-BiLSTM	0.83	0.89	0.86	0.76	0.78	0.77	0.71	0.67	0.69	0.77	0.78	0.77	0.72	0.77	0.74
Single task Pre-trained BERT	0.85	**0.92**	0.88	0.79	0.82	0.80	0.69	0.74	0.71	0.79	0.77	0.78	0.82	0.85	0.83
Single Task S-BERT-CNN-BiLSTM	0.85	0.89	0.88	0.80	0.87	0.83	0.71	0.75	0.73	0.76	0.86	0.8	0.79	0.89	0.83
Multi-Task S-BERT-CNN-BiLSTM	**0.93**	0.89	**0.91**	**0.81**	**0.89**	**0.85**	**0.79**	**0.85**	**0.82**	**0.82**	**0.92**	**0.87**	**0.86**	**0.92**	**0.89**

Table 3: Sample violation events picked up from News articles of different categories and are mapped across impacts. Note that the target organization names were masked by the token [ORGName] to maintain anonymity.

Classified Impacts	**Sample violation phrases picked up by the model**
Air pollution	a) The settlement addresses [ORGName]'s **failure to capture and control air emissions from storage vessels and to comply with associated inspection , record keeping and reporting requirements**. b) The alleged violations included **failure to manage and contain hazardous wastes; failure to comply with air emission limits; failure to comply with chemical accident prevention safety requirements; and failure to timely report use of certain toxic chemicals**.
Soil and natural resources	a) The case stems from several transformer spills at locations in Massachusetts and Connecticut, involving **improper manifesting of PCB remediation waste, improper storage of a PCB transformer, and improper disposal of PCBs** b) violations included discharges of pollutants primarily chlorides and sodium in excess of its permit, failure to properly monitor and maintain records, and **failure to adequately operate and maintain its wastewater treatment system.**

are detected correctly. The errors occur due to a portion of the sequence not detected correctly.

The primary reason for the poor performance of LLAMA-2 can be attributed due to two reasons: a) lack of environmental domain knowledge due to which critical domain concepts like, PCB remediation, PM2.5, PM10, bee harvesting etc. gets ignored. b) Unable to identify violation phrase boundaries. We observe that despite in most of the cases LLAMA-2 correctly identified the violation phrases, but the span of the phrases are either too long or too short. as a results of which outputs of the model get penalized. Similar observations were made for mistral 7B, however, since the mistral model is fine-tuned over the current dataset, problems related to domain concept mismatch were relatively less. However, the output word span still remains a challenge.

In terms of the active learning setup, we have observed the models to performs better at each round when new, clean human annotations are added. However, intelligently selecting the appropriate samples for active learning still remains a challenge. An area on which we need to explore in our future work.

3.1 Clustering violations

To derive additional insights, the environmental violations extracted from each set of articles were clustered. Embeddings for the sequences labeled as violation were created using the Universal Sentence Encoder (Cer et al., 2018). These vectors were then clustered using the K-means clustering algorithm (Aggarwal and Zhai, 2012), with cosine similarity as the underlying distance measure. We have used the popular Elbow method (Joshi and Nalwade, 2013) for selecting the optimal number of clusters, k, for each set.

Figure 1 depicts the different clusters obtained from environmental violations mined from the corpus, along with their percentage occurrences. The clusters reveal that many organizations are operating without necessary certifications required to ensure clean air and clean water at their premises.

4 Conclusion

In this paper, we have proposed computational models for extraction and curation of environmental incidents and violations from digitally published regulatory reports. A portion of this corpus has been manually annotated to train and evaluate a deep neural network architecture for automated extraction and curation of sustainability incidents and violations. Knowledge about sustainability events, violations, awards and penalties were used for the annotation task. The model is multi-tasking in nature. It simultaneously classifies a sentence as positive, negative or neutral and also labels portions of the sentence as incidents, violations or awards. The proposed multi-task network has been extensively evaluated with respect to some of the state of the art baseline models. We observed that for almost all the defined tasks the proposed model surpasses the baseline models.

References

2017. Kpmg survey of corporate responsibility reporting 2017.

2018. Esg incidents and value destruction: insights from the sustainalytics incidents integration study.

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Tushar Goel, Palak Jain, Ishan Verma, Lipika Dey, and Shubham Paliwal. 2020. Mining company sustainability reports to aid financial decision-making. In *AAAI-20 KDF-The AAAI-20 Workshop on Knowledge Discovery from Unstructured Data in Financial Services*.

Tian Guo. 2020. Esg2risk: A deep learning framework from esg news to stock volatility prediction. *Available at SSRN 3593885*.

Kalpana D Joshi and PS Nalwade. 2013. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing*, 2(7):219–223.

Jaeyoung Lee and Misuk Kim. 2023. Esg information extraction with cross-sectoral and multi-source adaptation based on domain-tuned language models. *Expert Systems with Applications*, 221:119726.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Christopher D. Manning, Bauer Surdeanu, Mihai, Finkel John, Bethard Jenny, Steven J., and David. McClosky. 2014. The stanford corenlp natural language processing toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Shinsuke Murakami and Shunya Muraoka. 2022. Exploring the potential of internet news for supply risk assessment of metals. *Sustainability*, 14(1).

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2020. Detecting esg topics using domain-specific language models and data augmentation approaches. *arXiv preprint arXiv:2010.08319*.

Stefan Pasch and Daniel Ehnes. 2022. Nlp for responsible finance: Fine-tuning transformer-based models for esg. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536. IEEE.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Amir Mohammad Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, 13(01):1450006.

Sayema Sultana, Norhayah Zulkifli, and Dalilawati Zainal. 2018. Environmental, social and governance (esg) and investment decision in bangladesh. *Sustainability*, 10(6):1831.

Indarawati Tarmuji, Ruhanita Maelah, and Nor Habibah Tarmuji. 2016. The impact of environmental, social and governance practices (esg) on economic performance: Evidence from esg score. *International Journal of Trade, Economics and Finance*, 7(3):67.

Patrick Velte. 2017. Does esg performance have an impact on financial performance? evidence from germany. *Journal of Global Responsibility*.

Chen Wan, Wenzhong Li, Wangxiang Ding, Zhijie Zhang, Qingning Lu, Lin Qian, Ji Xu, Jixiang Lu, Rongrong Cao, Baoliu Ye, et al. 2021. Multi-task sequence learning for performance prediction and kpi mining in database management system. *Information Sciences*, 568:1–12.

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2018. Multi-task learning for machine reading comprehension. *arXiv preprint arXiv:1809.06963*.

Bohyun Yoon, Jeong Hwan Lee, and Ryan Byun. 2018. Does esg performance enhance firm value? evidence from korea. *Sustainability*, 10(10):3635.

Changhong Zhao, Yu Guo, Jiahai Yuan, Mengya Wu, Daiyu Li, Yiou Zhou, and Jiangang Kang. 2018. Esg and corporate financial performance: Empirical evidence from china's listed power generation companies. *Sustainability*, 10(8):2607.